Grouping time series by pairwise measures of redundancy

D. Marinazzo[1], W. Liao[2], M. Pellicoro[3,4,5], and S. Stramaglia[3,4,5]
[1] *Laboratoire de Neurophysique et Neurophysiologie, Université Paris Descartes, Paris, France*

[2] *Key Laboratory for NeuroInformation of Ministry of Education, School of Life Sciences and Technology, University of Electronic Science and Technology of China, China*

[3] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy*

[4] *Dipartimento di Fisica, University of Bari, Italy and*
[5] *TIRES-Center of Innovative Technologies for Signal Detection and Processing, Università di Bari, Italy*

(Dated: June 25, 2010)

A novel approach is proposed to group redundant time series in the frame of causality. It assumes that (i) the dynamics of the system can be described using just a small number of characteristic modes, and that (ii) a pairwise measure of redundancy is sufficient to elicit the presence of correlated degrees of freedom. We show the application of the proposed approach on fMRI data from a resting human brain and gene expression profiles from HeLa cell culture.

PACS numbers: 05.45.Tp,87.19.L-

Over the last years the interaction structure of many complex systems has been mapped in terms of graphs, which can be characterized using tools of statistical physics [1]. Dynamical networks model physical and biological behavior in many applications; examples range from networks of neurons [2], Josephson junctions arrays [3] to genetic networks [4], protein interaction nets [5] and metabolic networks [6]. Synchronization in dynamical networks is influenced by the topology of the network [7]. The inference of dynamical networks is related to the estimation, from data, of the flow of information between variables. Two major approaches are commonly used to estimate the information flow between variables, transfer entropy [8] and Granger causality [9].

An important notion in information theory is the redundancy in a group of variables, formalized in [10] as a generalization of the mutual information. A formalism to recognize redundant and synergetic variables in neuronal ensembles has been proposed in [11] and generalized in [12]. Recently a quantitative definition to recognize redundancy and synergy in the frame of causality has been provided [13] and it has been shown that the maximization of the total causality, over all the possible partitions of variables, is connected to the detection of groups of redundant variables; the search over all the partitions is unfeasible but for small systems. We remark that the information theoretic treatments of groups of correlated degrees of freedom can reveal their functional roles in complex systems. The purpose of this work is to propose a simple approach to find groups of causally redundant variables (groups of variables sharing the same information about the future of the system), which can be applied also to large systems. The main assumption underlying our approach is that the essential features of the dynamics of the system under consideration are captured using just a small number of characteristic modes. Hence we use principal components analysis to obtain a compressed representation of the future state of the system. Then, we introduce a pairwise measure of the redundancy w.r.t. the prediction of the next configuration of the modes, thus obtaining a weighted graph. Finally, by maximizing the modularity [7], we find the natural modules of this weighted graph and identify them with the groups of redundant variables. In the following section we describe the method. In section II we describe the application of the method to fMRI data, and in section III to a gene expression data-set. Some conclusions are drawn in section IV.

I. METHOD

Let us consider n time series $\{x_i(t)\}_{i=1,\ldots,n}$; after a linear transformation, we may assume all the time series to be normalized and with zero mean. The lagged times series are denoted $X_i(t) = x_i(t-1)$. We make the hypothesis that the dynamics of the system under consideration may be described in terms of a few modes, and that these modes may be extracted by principal components analysis, as follows. Calling $\mathbf{x}$ the $n \times T$ matrix with elements $x_i(t)$, we denote $\{u_\alpha(t)\}_{\alpha=1,\ldots,n_\lambda}$ the (normalized) eigenvectors of the matrix $\mathbf{x}^\top\mathbf{x}$ corresponding to the largest n_λ eigenvalues. The T-dimensional vectors $u_\alpha(t)$ summarize the dynamics of the system; the lagged correlations of the system determine

to what extent the modes u may be predicted on the basis of the $X_i(t)$ variables.

Preliminarily, we select the variables which are significatively correlated with the modes u. For each i and each α we evaluate the probability $p_{i\alpha}$ that the correlation between X_i and u_α is due to chance, obtained by Student's t test. We compare $p_{i\alpha}$ with the 5% confidence level after Bonferroni correction (the threshold is $0.05/(n \times n_\lambda)$) and retain only those variables X_i which are significatively correlated with at least one mode. The variables thus selected will be denoted $\{Y_i(t)\}_{i=1,\dots,N}$, N being their cardinality.

The second step of the present approach is the introduction of a bivariate measure of redundancy, as follows. For each pair of variables Y_i and Y_j, we denote P_i the projector onto the one-dimensional space spanned by Y_i and P_j the projector onto the space corresponding to Y_j; P_{ij} is the projector onto the bi-dimensional space spanned by Y_i and Y_j. Then, we define:

$$c_{ij} = \sum_{\alpha=1}^{n_\lambda} \left(||P_i u_\alpha||^2 + ||P_j u_\alpha||^2 - ||P_{ij} u_\alpha||^2 \right); \tag{1}$$

according to the discussion in [13], c_{ij} is positive (negative) if variables Y_i and Y_j are redundant (synergetic) w.r.t. the prediction of the future of the system. In other words, if Y_i and Y_j share the same information about u, then c_{ij} is positive.

In the third step, the matrix c_{ij} is used to construct a weighted graph of N nodes, the weight of each link measuring the degree of redundancy between the two variables connected by that link. By maximization of the modularity [14], the number of modules, as well as their content, is extracted from the weighted graph. Each module is recognized as a group of variables sharing the same information about the future of the system.

As an example we report the following example. Let us consider the following autoregressive system:

$$\begin{aligned} \psi_t &= 0.6\eta_{t-1} + 0.1\xi_t^1 \\ \eta_t &= 0.6\psi_{t-1} + 0.1\xi_t^2, \end{aligned} \tag{2}$$

where ξ are i.i.d. unit variance Gaussian variables. By construction, ψ is caused by η and viceversa. A system of 50 time series is constructed as follows. For $i = 1, \dots, 10$:

$$\begin{aligned} x_i(t) &= \psi_t + 0.2\rho_t^i, \\ x_{10+i}(t) &= \eta_t + 0.2\rho_t^{10+i}, \\ x_{20+i}(t) &= \xi_t^3 + 0.2\rho_t^{20+i}, \\ x_{30+i}(t) &= \xi_t^4 + 0.2\rho_t^{30+i}, \\ x_{40+i}(t) &= \rho_t^{40+i}, \end{aligned} \tag{3}$$

where ρ and ξ are i.i.d. unit variance Gaussian variables. Starting from a random initial configuration, the above equations are iterated and, after discarding the initial transient regime, n_s consecutive samples of the system are stored for further analysis. Note that the first ten variables share the same information corresponding to ψ, whilst the second ten variables share the information of η. The variables x_i, with $i = 21, \dots, 30$, form a group of variables with correlations at equal times, similarly to the group of variables with $i = 31, \dots, 40$. The variables x_i, with $i = 41, \dots, 50$, correspond to pure noise. In figure (1) the equal-times correlations of the system are depicted, for a typical case with $n_s = 500$, showing four groups of correlated variables. We perform the principal components analysis and retain a variable number n_λ of modes for the analysis.

In figure (2), top-left, we depict N, the number of selected variables, as a function of n_λ. For $n_\lambda = 3, 4, 5$, twenty variables (x_i with $i = 1, \dots, 20$) are selected; nineteen variables for $n_\lambda = 1, 2, 6, 7, 8$. Then, for each value of n_λ, the quantities c_{ij} are evaluated. We find that, in this example, the matrix c_{ij} is non-negative and can be treated as a weighted graph.

In figure (2), top-right, we plot the number of modules N_m we find by applying the method described in [14] to the matrix c_{ij}; the method correctly recognizes the two modules for each value of n_λ.

In figure (2), bottom-left, we plot a measure of the stability of the partition while going from $n_\lambda - 1$ to n_λ, defined as follows. We consider all the pairs of variables that are selected both in correspondence of $n_\lambda - 1$ and n_λ. The stability is one minus the fraction of pairs such that the variables are recognized to be in the same module in one instance and in different modules in the other instance. In this case the stability is always one; when the method is applied to real data, the stability curve may be helpful to fix the optimal number of modes n_λ.

Finally, in figure (2), bottom-right, the eigenvalues of the matrix $x^\top x$ are depicted. In this case it is clear that the optimal number of modes is four.

We remark that a suitable number of samples is needed to obtain reliable results. In figure (3) we depict the number of selected variables, for this example, as a function of n_s for three choices of n_λ: it vanishes as the number of samples decreases.

II. MODULAR ORGANIZATION OF BRAIN ACTIVITY

The fMRI signal can be regarded as a proxy for the underlying neural activity. Remote regions of the brain do not operate in isolation and there is a growing interest in studying the interactions and connectivity patterns between these regions, which have been investigated by independent component analysis [15], principal components analysis [16] and other approaches. Temporal and spatial functional networks, corresponding to spontaneous brain activity in humans, were derived in [17] on the basis of the equal-time correlation matrix. Modularity in the resting state of the human brain has also been studied in [18–20]. The connectivity structure of brain networks extracted from spontaneous activity signals of healthy subjects and epileptic patients has been analyzed in [21, 22].

Here we consider fMRI data from a subject in resting conditions, with sampling frequency 1 Hz, and number of samples equal to 500. A prior brain atlas is utilized to parcellate the brain into ninety cortical and subcortical regions, and a single time series is associated to each region. All the ninety time series are then band-passed in the range 0.01-0.08 Hz.

In figure (4), top-left, we depict N, the number of selected variables, as a function of n_λ. For $n_\lambda > 3$, all the ninety regions are recognized as influencing the future of the system. For each value of n_λ, the quantities c_{ij} are evaluated. In figure (4), top-right, the number of modules N_m we find by applying the method described in [14] to the matrix c_{ij} is depicted; this plot suggests the presence of four modules for $4 < n_\lambda < 8$. These values are the most stable, as it is clear from figure (2), bottom-left, where we plot the measure of the stability of the partition while going from $n_\lambda - 1$ to n_λ. It follows that the optimal value of n_λ is four, corresponding to a graph structure with four modules and modularity equal to 0.3. We find that module 1 includes brain regions from ventral medial frontal cortices which are primarily specialized for anterior default mode network, module 2 is typical referred to as posterior default mode network, module 3 mainly corresponds to executive control network and module 4 refers to the subcortical network. In figure (4), bottom-right, the eigenvalues of the matrix $x^\top x$ are depicted.

It is worth comparing the histogram of the values of c_{ij}, in this example (figure (5)-bottom), with those corresponding to a random choice of the modes u (figure (5)-top). In the random case, the pairs of variables are either redundant or synergetic, and the typical values of c are very small. On the data set at hand, the magnitude of the values of c's is much higher and variables are mostly redundant; indeed the c's are negative (with small absolute value) only for a few pairs of regions. We remark that the presence of a few small and negative weights does not influence significantly the output from the modularity algorithm of [14]: the output does not change if all c's with absolute value less than a threshold are set to zero (the threshold being chosen so that all the elements of the resulting matrix are nonnegative).

Averaging the time series belonging to each module, we obtain four time series and we evaluate the causalities between them: the result is displayed in figure (6). It is interesting to observe that module 4 influences all the three other modules but is not influenced by them, it is an out-degree hub; this is consistent with the fact that it corresponds to subcortical brain. Another striking feature is the clear interdependencies between modules 2 and 3. The reliability of this pattern needs to be assessed on a large population of subjects.

III. HELA GENE EXPRESSION REGULATORY NETWORK

HeLa [23] is a famous cell culture, isolated from a human uterine cervical carcinoma in 1951. HeLa cells have somehow acquired cellular immortality, in that the normal mechanisms of programmed cell death after a certain number of divisions have somehow been switched off. We consider the HeLa cell gene expression data of [24]. Data corresponds to 94 genes and 48 time points, with an hour interval separating two successive readings (the HeLa cell cycle lasts 16 hours). The 94 genes were selected, from the full data set described in [25], on the basis of the association with cell cycle regulation and tumor development. This data has been also considered in [26]. The static correlation analysis between time series, which is the result of regulation mechanisms with time scales faster than the sampling rate, revealed a highly related network with the presence of two modules: the first module was recognized as corresponding to the regulatory network of the transcriptional factor NFkB [27], whilst the second module appeared to be orchestrated by transcriptional factors p53 and STAT3. Use of bivariate Granger causality, in [26], has put in evidence 19 causality relationships acting on the time scale of one hour, all involving genes playing some role in processes related to tumor development.

As stated in [28], fundamental patterns underlie gene expression profiles. This suggest the use of the proposed approach on gene expression time series. In figure (6) we describe the application of the proposed approach on the HeLa data-set. The stable partition corresponds to $n_\lambda = 4$ and consists of two modules of 9 and 7 genes (the modularity is 0.1). The first module is characterized by the transcriptional factor NFkB and consists of NFkB, MCP-1, ICAM-1, Bcl-XL, IAP, A20, c-myc, TSP1, and Mcl-1. The second module is related to the transcriptional factor JunB, known to be a regulator of life and death of cells [29], and consists of JunB, IL-6, IkappaBa, P21, Noxa, c-jun and NRBP. Averaging the time series belonging to each module, and evaluating the causality between the two time

series thus obtained, we obtain a relevant (0.145) causality of the first module on the second one. We note that all the sixteen genes selected by our approach were recognized as interacting in [24]; nine of them were involved also in the causalities described in [26]. It is not surprising that different methods, on the same data set, provide slightly different results: currently available data size and data quality make the reconstruction of gene regulatory networks from gene expression data a challenge. In figure (5)-bottom) we depict the histogram of the values of c on this data-set, with those corresponding to choosing randomly the modes u (figure (5)-top). On this data, the values of c that we obtain are significatively greater than those one finds in the random case, and all the pairs are redundant.

IV. CONCLUSIONS

Grouping redundant time series reveals their functional roles in complex systems. In the frame of causal approaches, grouping redundant time series may reflect directed influence of one group over another. In this work we have proposed a novel approach which assumes that (i) the dynamics of the system can be described using just a small number of characteristic modes, and that (ii) a pairwise measure of redundancy is sufficient to elicit the presence of correlated degrees of freedom. Grouping is provided by the identification of the modules of the weighted graph of redundancies. The method may be seen as an alternative to the analysis of [13], which can be applied also for large systems. We have shown the effectiveness of the proposed approach in two applications, the analysis of fMRI data and the analysis of gene expression data. In both these applications usually linear interactions are sought for, and only lags of 1 are considered, therefore here we limited to consider a linear pairwise measure of redundancy, and considered only lags of 1. The generalization of the pairwise redundancy measure, here introduced, to the nonlinear case and to lags of higher order is matter for further work, along the lines described in [13].

[1] A.L. Barabasi, *Linked: the new science of networks*. (Perseus Publishing, Cambridge Mass., 2002); S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[2] L.F. Abbott, C. van Vreeswijk, Phys. Rev. **E 48**, 1483 (1993).
[3] K. Wiesenfeld, Physica **B 222**, 315 (1996).
[4] T.S. Gardner, D. Bernardo, D. Lorenz, J.J. Collins, Science **301**, 102 (2003).
[5] C.L. Tucker, J.F. Gera, P. Uetz, Trends Cell Biol. **11**, 102 (2001).
[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature **407**, 651 (2000).
[7] M.E. Newman, Phys. Rev. Lett. **89**, 208701 (2002); S. Boccaletti, D.U. Hwang, M. Chavez, A. Amann, J. Kurths, and L.M. Pecora, Phys. Rev. **E 74**, 16102 (2006).
[8] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000)
[9] C.W.J. Granger, Econometrica **37**, 424 (1969); M. Dhamala, G. Rangarajan, M. Ding, Phys.Rev.Lett. **100**, 18701 (2008); D. Marinazzo, M. Pellicoro, S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).
[10] M. Palus, V. Albrecht, I Dvorak, Phys. Lett. A **175**, 203 (1993).
[11] E. Schneidman, W. Bialek, M.J. Berry, Journal of Neuroscience **23** 11539 (2003).
[12] L.M. Bettencourt, V. Gintautas, and M.I. Ham, Phys. Rev. Lett. **100**, 238701 (2008).
[13] L. Angelini, M. de Tommaso, D. Marinazzo, L. Nitti, M. Pellicoro, and S. Stramaglia, Phys. Rev. E **81**, 037201 (2010).
[14] M. Newman, Phys. Rev. E **74**, 036104 (2006); M. Newman, PNAS **23**, 8577 (2006).
[15] W. Liao et al., "Selective aberrant functional connectivity of resting state networks in social anxiety disorder", (2010). Neuroimage, doi: 10.1016/j.neuroimage.2010.1005.1010.
[16] G. Lohmann, et al. (2010) Eigenvector Centrality Mapping for Analyzing Connectivity Patterns in fMRI Data of the Human Brain. PLoS ONE 5(4): e10232. doi:10.1371/journal.pone.0010232
[17] Y. He et al., PLoS ONE **4**, e5226 (2009).
[18] L. Ferrarini et al., Human Brain Mapping **30**, 2220 (2009).
[19] D. Meunier et al., Neuroimage **44**, 715 (2009).
[20] M. van den Heuvel, R. Mandl, H.H. Pol, PLoS ONE **3**, e2001 (2008).
[21] M. Chavez, M. Valencia, V. Navarro, V. Latora, and J. Martinerie, Phys. Rev. Lett. **104**, 118701 (2010).
[22] W. Liao et al., PLoS One **5**, e8525 (2010).
[23] J.R. Masters, Nature Reviews Cancer **2**, 315 (2002).
[24] A. Fujita et al., BMC System Biology **1:39**, 1 (2007).
[25] M.L. Whitfield et al., Mol. Biol. Cell **13**, 1977 (2002).
[26] D. Marinazzo, M. Pellicoro and S. Stramaglia, Phys. Rev. E **77**, 056215 (2008).
[27] J. Inoue, J. Gohda, T. Akiyama, K. Semba, Cancer Sci. **98**, 268 (2007).
[28] N.S. Holter et al., PNAS **97**, 8409 (2000).
[29] E. Shaulian, and M. Karin, Nature Cell Biology **4**, E131 (2002).

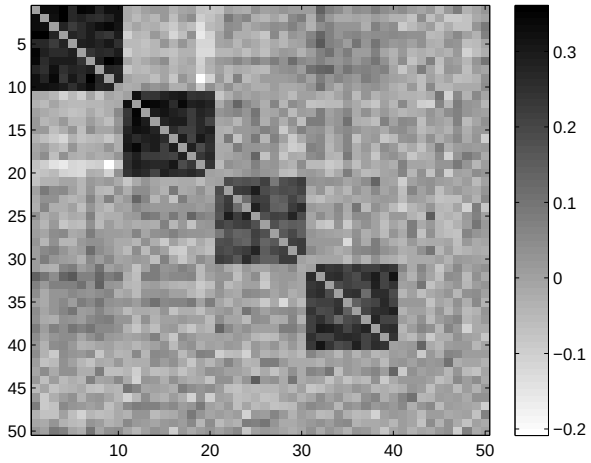


FIG. 1: The correlation matrix of the simulated example, showing four groups of variables correlated at equal times.

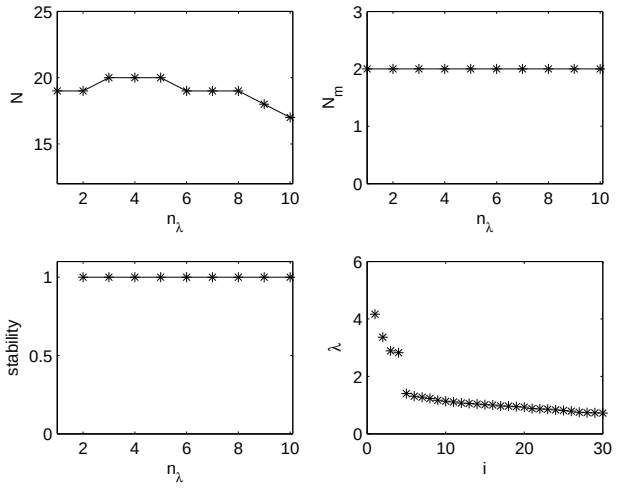


FIG. 2: (Top-left) Concerning the simulated example, the number of selected variables N is plotted versus n_λ, the number of modes. (Top-right) The number of modules, obtained by modularity maximization, of the matrix c_{ij}, whose elements measure the pairwise redundancy. (Bottom-left) The measure of the stability of the partition, going from $n_\lambda - 1$ to n_λ, is plotted versus n_λ. (Bottom-right) The eigenvalues of the matrix $x^\top x$ are depicted.

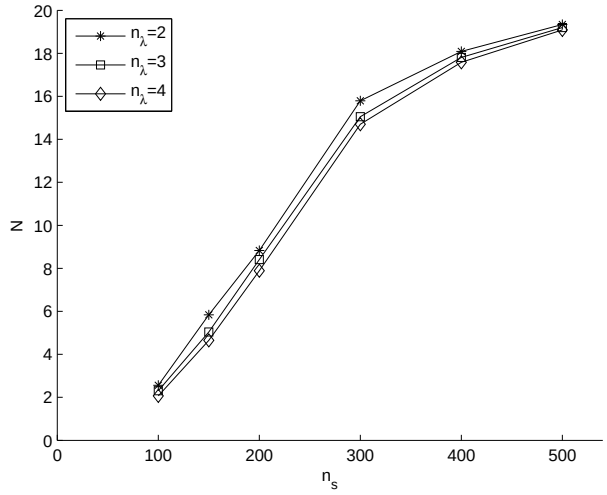


FIG. 3: The number of selected variables, for the simulated example, is depicted as a function of the number of samples n_s for $n_\lambda = 2, 3, 4$.

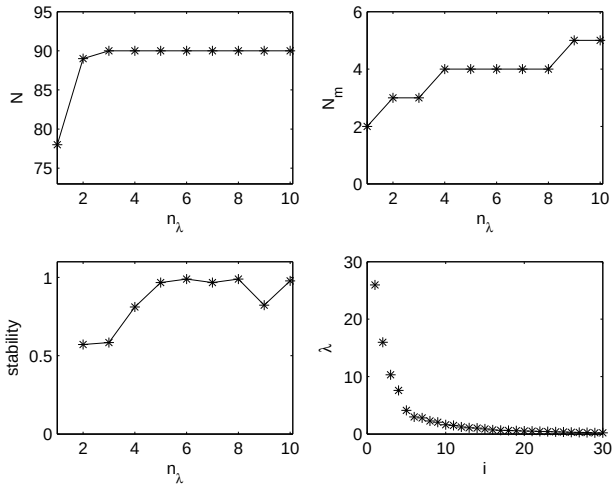


FIG. 4: (Top-left) Concerning the fMRI application, the number of selected regions N is plotted versus n_λ, the number of modes. (Top-right) The number of modules, obtained by modularity maximization, of the matrix c_{ij}, whose elements measure the pairwise redundancy. (Bottom-left) The measure of the stability of the partition, going from $n_\lambda - 1$ to n_λ, is plotted versus n_λ. (Bottom-right) The eigenvalues of the matrix $x^\top x$ are depicted.

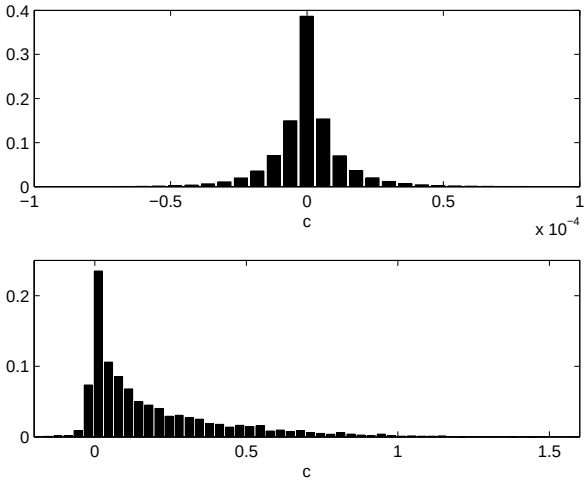


FIG. 5: The histogram of the values of the pairwise redundancy c_{ij}, in fMRI example (bottom), and choosing randomly the modes u (top)

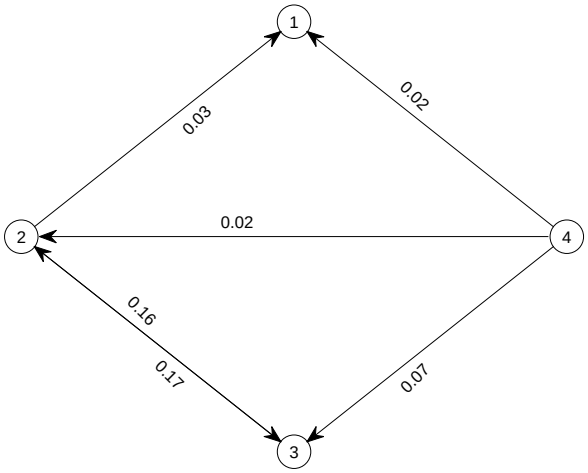


FIG. 6: The causalities between the four modules of the fMRI application.

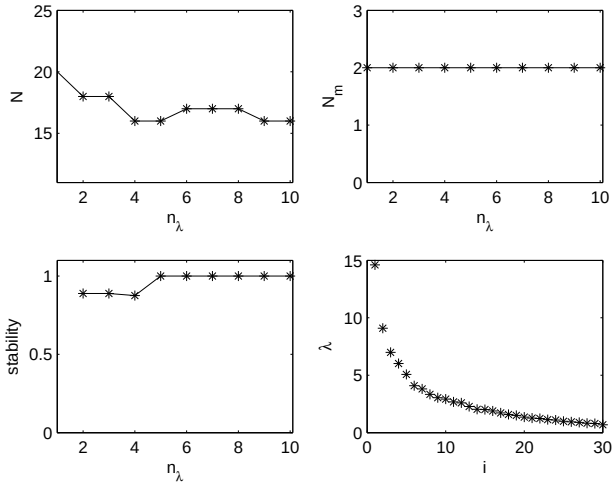


FIG. 7: (Top-left) Concerning the genetic application, the number of selected regions N is plotted versus n_λ, the number of modes. (Top-right) The number of modules, obtained by modularity maximization, of the matrix c_{ij}, whose elements measure the pairwise redundancy. (Bottom-left) The measure of the stability of the partition, going from $n_\lambda - 1$ to n_λ, is plotted versus n_λ. (Bottom-right) The eigenvalues of the matrix $x^\top x$ are depicted.

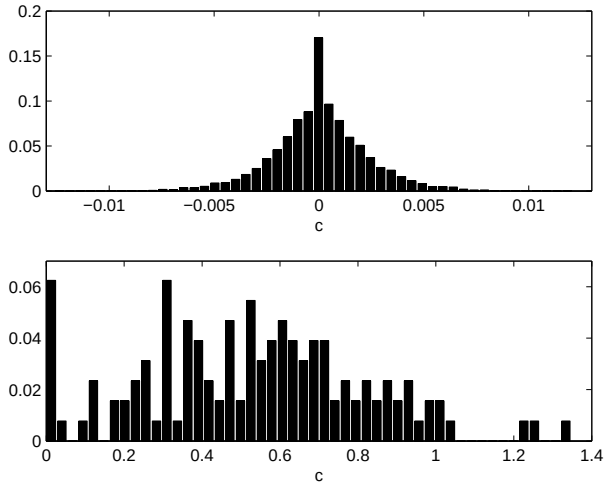
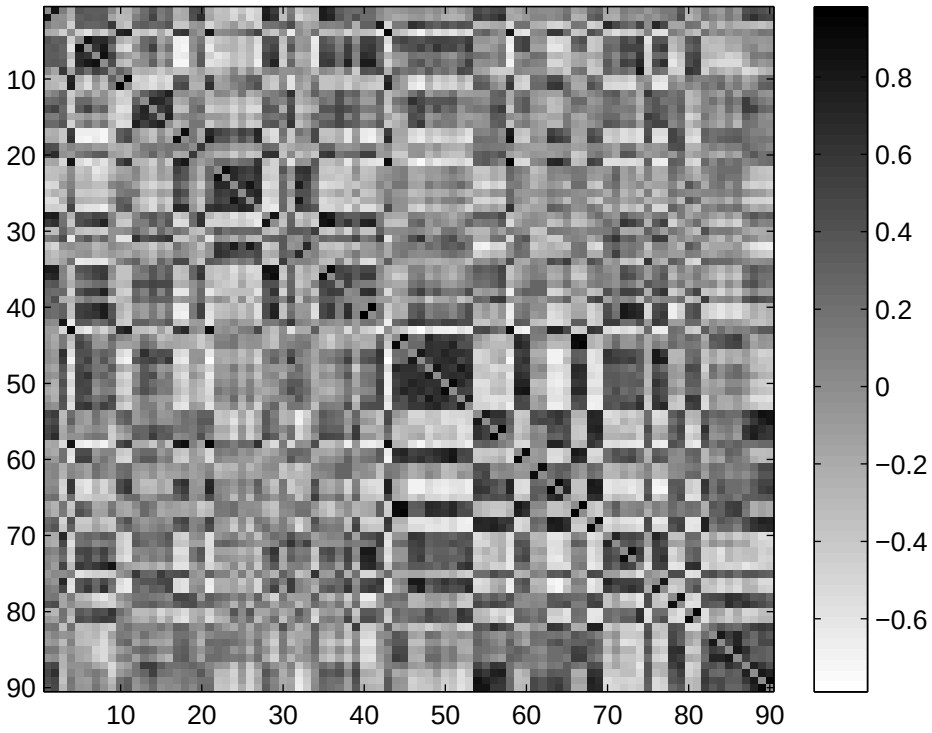


FIG. 8: The histogram of the values of the pairwise redundancy c_{ij}, in genetic application (bottom), and choosing randomly the modes u (top)


10
20
30
40
50
60
70
80
90
10
20
30
40
50
60
70
80
90
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6